Université de Liège
Faculté de Psychologie
et des Sciences de l'Education
Service de Technologie de l'Education
Dir. : Professeur D. LECLERCQ

Objective Assessment of Subjectivity : applying Confidence Marking to Partial Knowledge

to be presented at the Conference held by the EARLI SIG on « Assessment » at Maastricht on September 2000

D. Leclercq, V. Jans, F. Georges et J.-L. Gilles*

All from the University of Liège, Service de Technologie de l'Education
Contact : D. Leclercq@ulg.ac.be

Abstract

This paper reports the edumetrical developments (Carver, 1974) of De Finetti's (1965) claim : « Only subjective assessment can contribute to objective measurement of knowledge ».

During two decades (1960-1980), confidence marking, mainly applied to multiple choice questions, has inspired a series of research settings. It appeared to be a dead end because the experimental paradigm was polluted by big methodological flaws.

Adopting a correct methodology according to Shuford's (1966), Choppin's (1971) and De Finetti's (1965) principles enabled us to address a series of questions and to develop concepts and instruments to attempt to answer those questions, such as :

1. What is the human sensitivity (or limits of capacity) to estimate their partial knowledge ? A 20 % granularity ? A 10 % one ? A 5 % one ? Is there a "magical number 7" in this domain too ?
2. What is the degree of realism (tendency to overestimate or to underestimate) in terms of calibration graphics and scores ?
3. What is the relation between the expression of doubt and the information seeking behavior ?(When someone doubts, does he/she check information ?)
4. What is the "spectral" representation of a person's knowledge in a domain (from the highly confident errors to the highly confident correct answers ?
5. Can we observe "covert mental gains" or modifications in knowledge or opinions, that are not reflected in behavior, and what is the importance of those "moves" ?
6. Should we develop subjective analysis of questions in terms of new facility and discrimination indices ?

Those concepts have been applied during years, some in a repeated way on groups of 300 students, others in more "clinical" settings, and recently in a survey involving 4000 freshmen entering 8 Belgian universities (from French speaking community).

A. A preliminary issue : Why be afraid of CM ?

Why did the Confidence Marking boat sink between 1960 and 1980 ? Five methodological flaws contributed to make the whole endeavour fail (Leclercq, 1993, 213-218).

The first flaw consisted in a lack of rigor in the definition of the DCs (Degrees of Confidence) themselves, stating for instance the following instructions : "Are you sure ? Weakly sure ? Fairly sure ? Strongly sure ? etc." Those "verbal" instructions are far too vague and cannot be compared with reality, as suggested by Shuford *et al.* (1966) in their famous Psychometrika article "Admissible probability measurement procedures" and according to De Finetti's (1965, 111) sentence "It is only subjective probability that can give an objective meaning to every response and scoring method". Instructions must refer to a metric scale, with Confidence Degrees such as 0%, 20%, 40%, 60%, 80%, 100%. This is the last one we have experimented, among a series of others (Leclercq, 1983, 1993).

The second flaw was a lack of respect of decision theory in the definition of the scale of tariffs. Actually, tariffs, were created by "rules of thumb" (Lindley, 1971; Luce & Raiffa, 1966; Raiffa, 1970; Savage, 1951), or according to classical "correction for guessing formulas", that have been demonstrated by Choppin (1971,1974, 1975) as based on incorrect models of a student's mental activity. Consequently, some tariff scales encouraged students to lie about their confidence in order to maximise their score to each item, and consequently their score at the total of the test, most of the time without the students and the teachers being aware of this process. Tariffs compatible with decision theory have been described by Leclercq (1993, 214); Van Lenthe (1993, 132-145); Dirkzwager (1993, 146-166) and Shuford (1993, 76-98).

The third flaw was a lack of conceptualisation of the new score as a payment combining two measures. This flaw is revealed by the (wrongly stated) question *: "Are new (total) test scores (computed with new scales of tariffs taking confidence degrees into account) more valid and more reliable than classical ones (number of correct answers) ?"* Results from these experiments to check this point are confusing. Half of the studies find they are more valid and less reliable, whereas the other half of the studies find the contrary…without being able to explain these contradictory results. Actually, when a learner expresses (through Confidence Degrees) what he knows about what he knows, the teacher has a more valid view of this knowledge only if the student is realistic. If not, the increase of data brings more noise than information (to use Shannon's words). As a consequence, the Confidence Degrees should be used to estimate knowledge more subtlely only after the student's Realism has been proved as sufficiently high (a 80 level of realism for instance). In the same way, since Performance and Metacognition (realism in self-assessment) can be evaluated and computed separately, constituting 2 measures, the new total score may not be a measure itself, but the combination (in proportions announced in advance) of these two different measures.

The fourth flaw was the lack of feedback about realism, since the trainers did not compute realism indices neither drew the calibration graphs, whereas those indices can be easily developed on the basis of Lichtenstein *et al.* (1997)'s principles, themselves grounded in the works of Brier (1950), Adams & Adams (1961), Oskamp (1962), Murphy (1972, 73, 74). We have called the difference between the average DC and the Objective Success Rate at the test, the "Error of Centration index" (when it is negative, it means underestimation , when it is positive, it indicates overestimation and when its value is 0, it means perfect centration). The "Internal Coherence index" is the correlation between confidence degrees and success rates and indicates how far the student is coherent with him/her self. The "Acuity – Subtlety index" is the standard deviation of the Objective Success Rates of the various Confidence Degrees. The "Realism index" (the formulas and norms of which can be found in Leclercq (1993, 127-130) expresses the proximity of the Observed Success Rates (OSRs) to the announced ones , i.e. the Confidence Degrees, or Predicted Success Rates (PSRs).

The fifth flaw was the lack of students training. Usually they are not familiar with the procedure. They never had before the opportunity to estimate their partial knowledge, to face their calibration curve, to observe their evolutions in Realism or Coherence indices from

successive tests, etc. Illustrative data of evolution due to training are provided in Leclercq (1993, 129).

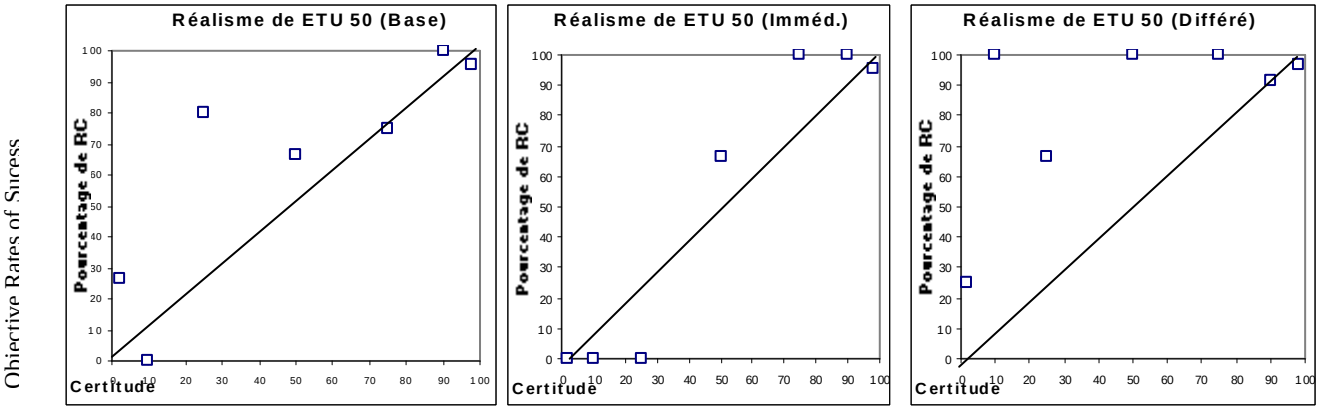
B. <u>Conceptual and methodological developments in Confidence Marking</u>

1. What is the human sensitivity of human beings in estimating their confidence degrees to make their answers more subtle ?

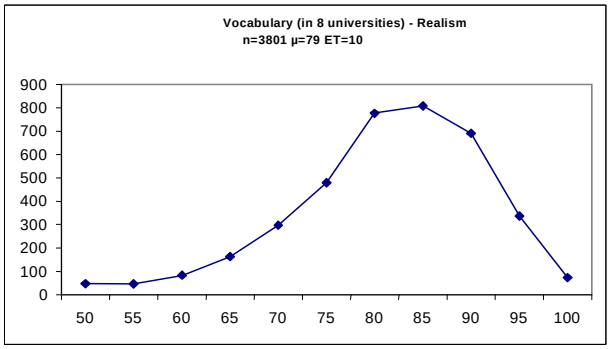
We have developed a "Confidence Guessing Game" (Leclercq, 1993, 121-126)[1] to study this systematically. It came out that adults' sensitivity or acuity or granularity is better in extreme portions of the probability axis (close to 0 % and close to 100 %) than in the centre (close to 50 %). We have explained (Leclercq, 1993, 125) why on the basis of Edward's theory (1967). Confidence Sensitivity can hardly exceed 7 portions on the probability axis, so that currently we recommend the following instructions : "In addition to your answer, provide a confidence degree among the 6 following ones : 0%, 20%, 40%, 60%, 80%, 100%." This will determine 6 categories of responses and for each of them we will be able to compute a Rate Of Success (ROS). It is obvious that if a student has provided 10 answers with the 60% confidence degree, we expect 6 of his answers to be correct out of 10 (i.e. a 60% ROS). If it is so, the student is "realistic" or "well calibrated" for that (60 %) Confidence Degree (CD).

2. What is the degree of realism students can achieve ?

A classical way of displaying realism is to draw the calibration diagram. Here is an example of 3 calibration diagrams of the same student (ETU 50) passing the same test at a pretest, then as an immediate post-test and differed post-test. His tendency to underestimate appears in the base (pretest) diagram and in the differed (post-test) one (Jans, 2000).



The formula for computing the realism index is : **Realism = 100 – MAEC**.
Where MAEC is the Mean Absolute[2] Error of Confidence. Here is the distribution of about 4000 freshmen entering in first year (in 8 universities) answering a 45 questions a Vocabulary test (resulting in 180.000 answers and 180.000 confidence degrees) in the MOHICAN testing (Leclercq and Georges, 2000).



[1] inspired by the Shannon Guessing Game (Attneave, 1959).

[2] i.e. in Absolute value.

3. What is the relation between the expression of doubt and consulting personal notes?

The instructions of an experimental design

Fifty graduate students of the university of Liège were given (orally and in written format) the following instructions :

"We want to study your note taking behavior in order to help further consultation of a hypermedia courseware (about 200 pages), just as one usually does when reading reference books (inserting pieces of paper between 2 pages, folding corners, writing signs in margins, etc.) or when viewing video (noting number count and key words to help remembering sequences). The successive events will be as follows.

1. *Today you will explore the content, using freely the possible itineraries and taking notes (in a dialog window box, with the help of the keyboard), essentially the page (screen) number and any clue you give to yourself to access it further on.*
2. *In two weeks, you will be asked to answer a 15 questions MCQ test. This will constitute Post-Test 1.*
3. *Your electronic notes will be made available to you in paper format, and you will be allowed to ask to see again 15 pages (screens) of your choice, knowing that the same test will be given to you and that you can change your answers. This is the Consultation phase.*
4. *A copy of your Post-Test 1, i.e. your answers and Confidence Degrees, will be given back to you and you will have the opportunity to maintain or to change them."* This constitutes Post-test 2.

The Test : SIG MCQs and Confidence Degrees

The test contains 15 MCQs, 8 of which been usual (the correct answer is one of the printed alternatives), the 7 others been "General Implicit Solution" (see Leclercq et al., 1993 b), i.e. either code 6 (None is correct), 7 (The Totality of them is all correct), 8 (Missing data to decide) or 9 (An Absurdity in the stem makes the whole question meaningless).

In addition, the students had to provide a Confidence Degree for each of their questions, on a 6 levels scale. Tariffs are computed according to decision theory so that students are interested in telling the truth (express their subjectively estimated confidence without bias).

Results

Consulting annotated screens improves mean number of correct answers for 43 students out of 50 (86%). Screens have been consulted in 35% of cases for a correct answer (on Pre-test) and 65% for a incorrect answer.

Screen consultation and lack of confidence

The relation between the confidence degree (on Pre-test) and the frequency of consulting screen at post-tests is as follows (Leclercq & Boskin, 1989).

When the Confidence Degree at Post-tests was ….	0%	20%	40%	60%	80%	100%
…the rate of consulting the corresponding page was	61%	52%	63%	58%	48%	26%

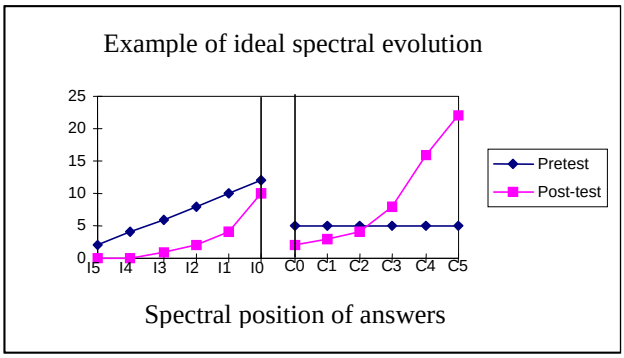
The more the students doubt, the more they check. The same kind of results have been observed by Jans (1999). These results support Descartes' view (1636) that "*doubt is the incentive of knowledge*" (pp. 126-127 in the 1952 edition) : the consulting behavior is explained by subjective reasons, not by the "objective" state of our knowledge. In the domain of health, Rosenstock (1973) has also demonstrated (in his "Health Belief Model") that behavior is driven by beliefs (of gravity, vulnerability, detectability and curability) about a disease more than by official information on this disease. This general principle is of major importance in learning strategies and in metacognition (Brown, 1978).

4. What is the "spectral" representation of a person's knowledge ?

JANS (1999) has suggested a spectral way of representing the continuum of responses, consisting of 2 hemispectra : the left one (incorrect answers) and the right one (correct answers). For each of those two hemispectra, it is possible to compute the skewness of distribution curve by the following formula :

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

For instance, for the following example (the vertical axis presenting the frequencies)

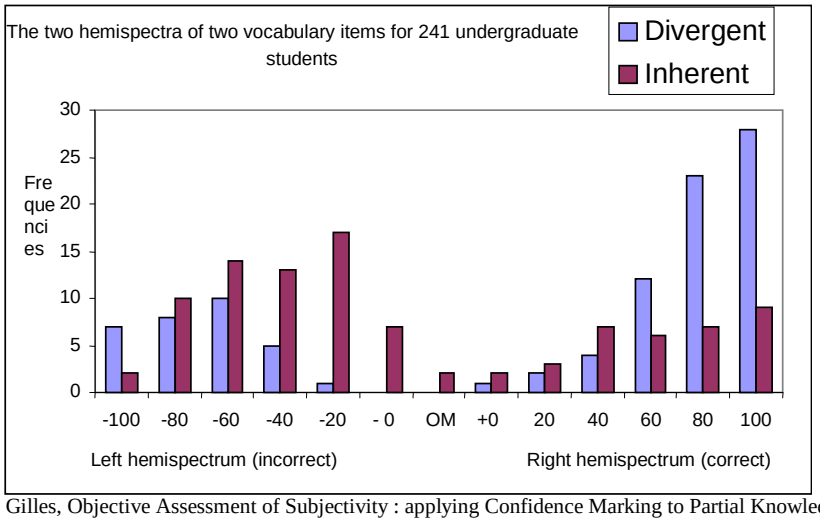


the **skewness** index values of the hemispectra are :

	Incorrect answers (left spectrum)	Correct answers (right spectrum)
Pretest	-0.49	0
Post-test	-2.03	-1.15

Contrarily to psychometrics, where the ideal curve is a Gauss shaped one, in edumetrics (Carver, 1974), the ideal curve is a J shaped one. So, the more left sided asymmetric the curve, the best; the more negative the skewness index, the best.

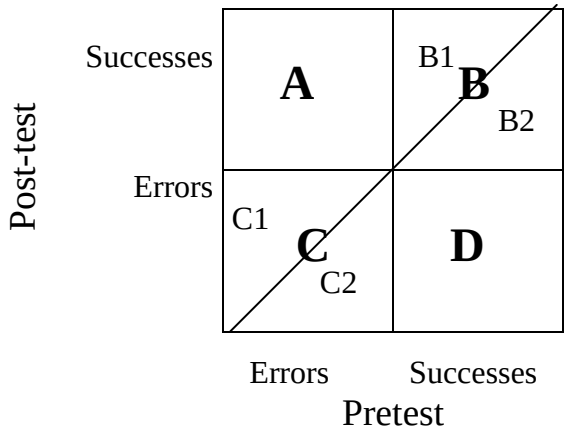
Here are the two hemispectra of the answers to two items out of a 45 questions vocabulary test (the words "inherent" and "divergent") given by 241 students entering a Faculty. Of course, the global hemispectra for all the (45) questions have also been established, as well as the merged hemispectra for the 4000 students from 8 universities who have passed this same test in October 1999.



It can been seen that for the item "Divergent", the right hemispectrum (the correct answers) is a perfect J shape but the left one (the incorrect answer) is far from a J shape. For the item "inherent", the right hemispectrum is less satisfactory but the left one is better than for the other item.

5. Can we observe "covert mental gains or modifications" ?

In the cognitive domain as well as in the motivational one, changes in mind are not always translated or expressed in changes in behaviors. Thanks to Confidence Degrees, this could be observed in two recent studies. In the **cognitive domain**, Jans (1999) collected the answers ans Confidence Degrees of several students passing a 100 open ended questions test on english vocabulary twice, as a pre-test and as a post-test (after having the opportunity of using a hypermedia courseware on the english language). In the **domain of opinions**, Rommes and Leclercq (1997) organized an animation where 23 students in educational psychology had to suggest the professor's best answer to the disruptive behavior demonstrated by a student in a classroom. This experiment was based on a real case according to the Programmed Case Method (Vandenbrande, 1994; Leclercq & Vandenbrande, 1997; Leclercq et al., 1998). The 23 students had to predict the disraptor's behavior twice : before and after a debate. The meaning of the four squares (A, B, C, D) hereafter are the following ones :



A = Error at pre-test and success in post-test
B = Success in pre and post-test

C = Error in pre and post-test
D = Success at pre-test and error in post-test

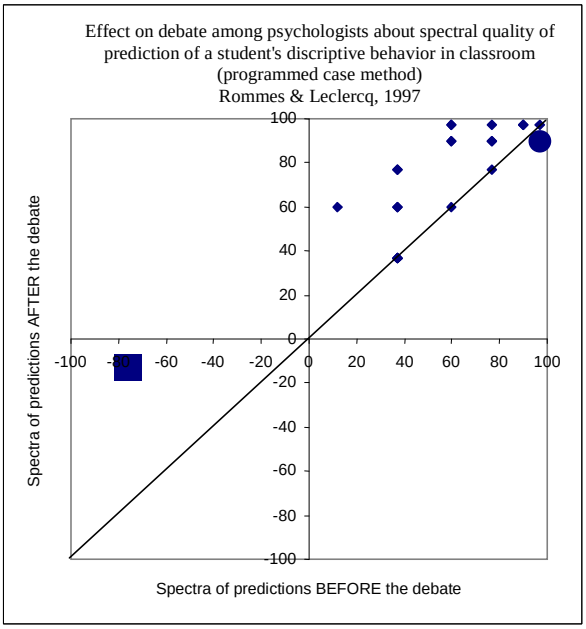
Objective status quo are represented by squares B and C. Objective improvement is represented by square A. Objective deterioration is represented by square D. Nevertheless, if squares C and B are each split into parts 1 and 2 (whether it is over or under the diagonal), subjective status quo is represented only by the points on the diagonal line. Subjective improvements are constituted by A + B1 + C1. Subjective deteriorations are constituted by D + B2 + C2.

	In Jans' case (1 student, 100 questions)	
	Objective	**Subjective**
Status quo	76	49
Improvement	23	47
Losses	1	4

In Rommes'case (23 students, 1 question)	
Objective	**Subjective**
23	4
0	18
0	1

In Rommes'case, remaining at the objective level of observation would come up with the conclusion that the debate had no impact at all (almost no gain, no loss) and should be discarded in further experiments, whereas subjective analysis shows (see diagram below) an improvement among a large majority of students.

On the graphic, it appears that only one person (the big circle) gave a "worse" answer after the debate. It is even more obvious from visual inspection that the improvement would have been underestimated if Confidence Degrees had not been collected. The single "0-1" (loss) of the experiment is figured by a big square.



6. Should we develop subjective spectral item analysis ?

Leclercq (2000, in press) has suggested to compute, in addition to the (objective) facility index (OF), i.e., the rate of success (or of correct answers) to an item,

- the SF (Subjective Facility), i.e. the average confidence degree for all answers
- the SFCA (Subjective Facility of Correct Answers), averaged over the students
- the DM(Degree of Mastery), computed by DM = OF x SFCA

Gilles (1999, 19-30) has suggested to compute Spectral discrimination indices, i.e. spectral point biserial correlations; his formula is as follows for question a :

$$\text{Spectral rpbis a} = \frac{\text{CMRC - CMRI}}{\text{SD}} \sqrt{pq}$$

where

CMRC = Mean Confidence for Correct Responses
CMRI = Mean Confidence for Incorrect Responses
p and q have the same meaning as in the classical point biserial correlation formula, i.e. p = proportion of correct answers and q = 1-p.
SD is the Standard Deviation of all the confidence degrees given to this question (regardless of the response being correct or incorrect).

He also computes (p.27) the “Turbo” Spectral discrimination index, i.e., the same formula applied only on the students who demonstrate a good level of realism (for instance with a realism > 80%. He is investigating the fecundity (compared with the classical point biserial one) of these new discrimination indices in terms of detection of inappropriate alternatives in the MCQs.

Conclusion

Everybody acknowledges the importance of metacognitive skills and mathetic[3] competencies for nowadays learners. Self assessment is only an aspect of them and Confidence Degrees are only one way among others to address the issue. We hope that the first part of this article has demonstrated that some restrictions that are legitimately associated with this technique are carefully taken into account, that there are valid and reliable ways to use Confidence Degrees, and that the second part of the article has demonstrated that this technique offers the potential for new and fecund approaches to old problems. We have decided not to enter the debate of the definition of competency, since place was lacking here and we wanted to focus on technical aspects. Nevertheless, we are confident that this approach can bring its special light in the old debate so well stated by T. S. Eliott :

> "Where is information we lost in data ?
> Where is knowledge we lost in information ?
> Where is wiseness we lost in knowledge ?"

[3] Word coined by Gilbert (1962) from the greek word μανθανω (I learn), to designate « in relation to learning ».

REFERENCES

Adams, J.K. & Adams, P.A. (1961), Realism of confidence judgments, Psychological Review, 68, 33-45.

Attneave, F. (1959), Application of Information Theory to Psychology, New-York : Holt, Rinehart & Winston.

Brown, A.L. (1978), Knowing when, where and how to remember : a problem of metacognition, Glaser, R. (1978), Advances in instructional psychology, vol. 1, Hillsdale : N.J., Lawrence Erlbaum Associates.

Bruno, J. (1993). Using Testing to Provide Feedback to Support Instruction : A Reexamination of the Role of Assessment Organizations, in: Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 190-209.

Brier, G.W. (1950), Verification of forecasts expressed in terms of probability, Montly Weather Review, 75, 1-3.

Carver, R.P. (1974), Two dimensions of tests: psychometric and edumetric, American Psychology, 29, 512-518.

Choppin, B. (1970), An IEA Study of Guessing. A Proposal, Stockholm : International Association for the Evaluation of Educational Achievement, unpublished memorandum, IEA/TR/9.

Choppin, B. (1974), The Correction for Guessing on Objective Tests, Stockholm : International Association for the Evaluation of Educational Achievement.

Choppin, B. (1975), Guessing the Answer on Objective Tests, Brit. Journal of Educational Psychology, 45, 206-213.

Cronbach, L. & Snow, R. (1977), Aptitudes and Instructional Methods, A handbook for research on interactions, New York : Wiley.

De Finettti, B. (1962), Does it make sense to speak of good probability appraisers ? in GOOD, I.J. (Ed.), The Scientist Speculates, New York : Basic Books, 357-364.

De Finetti, B. (1963), La décision et les probabilités, Bucarest : Revue des Mathématiques pures et appliquées, 405-413.

De Finetti, B. (1965), Methods for discriminating levels of partial knowledge concerning a test item, British Journal of Mathematical and Statistical Psychology , 18, 87-123.

De Finetti, B. (1970), Logical Foundations and Measurement of Subjective Probability, Acta Psychologica, 34, 129-145.

Denis, B. & Leclercq, D. (1994), The Fundamental Instructional Designs (ID) and their associated problems, in: Jowyck, J. & Elen, J. (Eds), Modelling ID Research, proceedings of the SIG EARLI Workshop on Instructional Design, Leuven, 67-85.

Descartes, R. (1636), Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences.

Dirkzwager, A. (1993), A Computer Environment to Develop Valid and Realistic Predictions and Self-Assessment of Knowledge with Personal Probabilities, in: Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 146-166.

Fabre, J.M. (1993), Subjective Uncertainly and the Structure of the Set of all Possible Events, in: LECLERCQ D. & BRUNO J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 99-113.

Gilbert, J.-F.(1962), Mathetics : The Technology of Education, Journal of Mathetics, 1, 7-74.

Gilles, J.-L. (1996), Utilisation des degrés de certitude et normes de réalisme en situation d'examen et d'auto-estimation à la Faculté de Psychologie et des Sciences de l'Education de l'Université de Liège, Actes du colloque de l'ADMEE-EUROPE « Dix années de travaux de recherche en évaluation », septembre 1996, Université Pierre Mendes France à Grenoble.

Gilles, J.L. (1999), Apports des mesures métacognitives lors d'un test de compréhension d'un article scientifique, in C. Depover & B. Noël (Eds), Approches plurielles de l'évaluation des compétences et des processus cognitifs, Proceedings of the 12th ADMEE Conference, Mons; UMH-FUCAM, 19-30.

Hunt, D. (1993), Human Self-Assessment : Theory and Application to Learning and Testing, in: Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 177-189.

Jans, V. & Leclercq, D. (1997), Metacognitive realism : a cognitive style or a learning strategy ?, Educational Psychology, vol. 17, 1 & 2, march.

Jans, V. (1999), L'autoévaluation de performances simples et complexes par les étudiants universitaires : description et résultats d'une expérience, in C. Depover & B. Noël (Eds), Approches plurielles de l'évaluation des compétences et des processus cognitifs, Proceedings of the 12th ADMEE Conference, Mons; UMH-FUCAM, 61-76.

Jans, V. (2000), Confrontations instrumentées et dialectiques des jugements auto/alloévaluatifs, Ph. D. Dissertation in Education, Liège University.

Kirk, R.E. (1996), Practical significance : a concept whose time has come, Educational & Psychological Measurement, vol. 56, n° 5, oct. 96, 746-759.

Laveault, D. & Gregoire, J. (1997), Introduction aux théories en sciences humaines, Bruxelles : De Boeck.

Leclercq, D. (1983), Confidence marking, its use in testing, Evaluation in Education Series, Oxford : Pergamon Press.

Leclercq, D. (1993), Validity, Reliability, and Acuity of Self-Assessment in Educational Testing, in: Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 114-131.

Leclercq, D. & Boskin, A. (1990), Note Taking behaviors studied with the help of hypermedia, in ESTES, HEENE et LECLERCQ (Eds). Proceedings of the 7th International Conference on Technology and Education, Brussels.

Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag.

Leclercq, D. & Gilles, J.-L. (1993), Hypermédia : Teaching through Assessment, in: Leclercq D. & Bruno J. (1993a), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 31-48.

Leclercq, D., Boxus, E., de Brogniez, Ph., Wuidar, H., Lambert, F., (1993b), The TASTE Approach : General Implicit Solutions in MCQs, Confidence Marking, Open Books Exams and Interactive Testing, in :Leclercq, D. and Bruno, J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag.

Leclercq, D. & Gilles, J.-L. (1994), GUESS, Un logiciel pour s'entraîner à l'auto-estimation de sa compétence cognitive, in: Weber, J. & Dumont, B. (1994), QCM et Questionnaires Fermés, actes du 3° colloque international de l'ESIEE, Marne-la-Vallée, décembre, 137-158.

Leclercq, D. & Vandenbrande, L. (1997), Une méthode pour la formation universitaire clinique en criminologie, psychologie et éducation : les cas programmés, Communication présentée au 15e Colloque AIPU, Liège.

Leclercq, D., Georges, F., Gilles, J.-L., Reggers, Th., Rommes, O. (1998), Interactive Multimedia Programmed Biographies (IMPB) : a new method for clinical training, Proceedings of at the BITE (Bringing Information Technology for Education) Conference, Maastricht, 25-27 March 1998, 406-417

Leclercq, D., Georges & Gilles, The MOHICAN project results, Bruxelles, CIUF : in progress.

Leclercq, D., Gilles, J.-L. & Georges, F. (2000), The MOHICAN study, Historical Monitoring of Candidatures, Bruxelles, CIUF, in progress.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1977), Calibration of probabilities : The state of the art, in Jungermann & DeZeeuw (Eds), Decision making and change in human affairs. Proceedings of the 5th SPUDM Conference, Darmstadt : Reidel Publishing Company, 275-324.

Lindley, D.V. (1971), Making decisions, London : Wiley.

Luce, R.D., Raiffa, H. (1966), Games and Decision, New-York : Wiley.

Mc Guigan, F.J. (1967), The G Statistics, An Index of Amount Learned, NSPI Journal, 69, 14-16.

Murphy, A.H. (1974), A sample skill score for probability forecasts. Montly Weather Review, 102, 48-55.

Nielsen, J. (1990), Hypertext and Hypermedia, San Diego : Academic Press.

Noel, B. (1991), La métacognition, Bruxelles : De Boeck.

Oskamp, S. (1962), The relationship of clinical experience and training methods to several criteria of clinical prediction, Psychological Monographs, 76.

Raiffa, H. (1970), Decision Analysis, Introductory Lectures on Choice under Uncertainty, New-York : Addison-Weslay.

Rosenstock, I.M. (1974), Historical origins of the Health Belief Model, Health Education Monographies, 2, 328-335.

Savage, L.J. (1951), The Foundations of Statistics, New-York : Wiley.

Shuford, E., Albert, A. et Massengil, N.E. (1966), Admissible Probability Measurement Procedures, Psychometrika, 31, 125-145.

Suford, E.H. (1967), How to Shorten a Test and Increase its Reliability and Validity, Lexington, Shuford-Massengil Corporation, technical report SCMR-8.

Shuford, E. (1969), Systems of Confidence Weighting, Theory and Practice, Los Angeles : Annual Meeting of the American Educational Research Association.

Shuford, E.H. (1993), In Pursuit of the Fallacy : Resurrecting the Penalty, in: Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series, F 112, Berlin : Springer Verlag, 76-98.

Van Lenthe, J. (1993), The Development and Evaluation of ELI, an Interactive Elicitation Technique for Subjective Probability Distributions, in: Leclercq D. & Bruno J. (1993), Item Banking : Interactive Testing and Self-Assessment, NATO ASI Series F112, Berlin : Springer Verlag, 132-145.

Vanden Brande, L. (1994), Training in Getting to know a Person, Ph D thesis in Psychology, University of Liège.